A Cascaded Broadcast News Highlighter

Heidi Christensen, Yoshihiko Gotoh, and Steve Renals. A cascaded broadcast news highlighter. IEEE Transactions on Audio, Speech and Language Processing, 16:151-161, 2008.

A Cascaded Broadcast News Highlighter

Heidi Christensen, Yoshihiko Gotoh, and Steve Renals, *Member, IEEE,*

***Abstract*— This paper presents a fully automatic news skimming system which takes a broadcast news audio stream and provides the user with the segmented, structured and highlighted transcript. This constitutes a system with three different, cascading stages: converting the audio stream to text using an automatic speech recogniser, segmenting into utterances and stories and finally determining which utterance should be highlighted using a saliency score. Each stage must operate on the erroneous output from the previous stage in the system; an effect which is naturally amplified as the data progresses through the processing stages. We present a large corpus of transcribed broadcast news data enabling us to investigate to which degree information worth highlighting survives this cascading of processes. Both extrinsic and intrinsic experimental results indicate that mistakes in the story boundary detection has a strong impact on the quality of highlights, whereas erroneous utterance boundaries cause only minor problems. Further, the difference in transcription quality does not affect the overall performance greatly.**

***Index Terms*— statistical modelling, spoken language processing, speech understanding, information extraction**

I. Introduction

Ever increasing amounts of spoken data are becoming available in digital form: through media such as TV and radio broadcasts and podcasts, and in the form of recordings of meetings, lectures, presentations, and personal interactions such as telephone conversations. The value of such archived data will be increased by the availability of tools that facilitate access by enabling efficient skimming and exploration. Compared to text, people often find it difficult to quickly skim unprocessed spoken audio; in this paper we discuss an approach to segment and structure such data in terms of topics (or stories), utterances and highlights.

A useful speech highlighter will operate in a fully automatic fashion on the original (or near-to-original) audio signal, eliminating the need for manual annotation. A general cascaded framework for such a system builds on technologies developed for processing written documents, in which the initial step is to transcribe the audio signal, followed by a segmentation component that provide utterance and story boundaries.

This paper presents a news skimming system — a fully automatic system for providing highlights of broadcast news programs. The system takes a broadcast news audio stream and provides the user with the segmented, structured and *highlighted* transcript. This constitutes a system with three different stages: initially the audio stream is converted to text using an automatic speech recogniser (ASR), it is subsequently segmented into smaller, coherent units (stories and utterances), and at the final stage a highlight utterance is extracted for each story, based on the utterance *saliency score*, which measures the information relevance of each individual utterance.

ASR, story and utterance segmentation, and highlighting are not solved problems. Each stage must operate on the erroneous output from the previous stage in the system; an effect which is naturally amplified as the data progresses through the processing stages. We are interested in establishing to which degree usable information survives this cascading of stages and is ultimately available for the user to review.

Broadcast news data is highly suitable for this study: it is of general interest due to its global availability as described above. Further it has for some years been the focus of intensive research in the automatic speech recognition community resulting in relatively high accuracy when producing text transcripts of the spoken material. The same applies to segmentation approaches, thus making these technologies ripe for incorporating into larger-scale systems and lending themselves to being analysed in a more holistic manner. However, very little is known about the effects of the interaction between each individual stage.

Establishing a news skimming system touches on a number of different research disciplines concerned with processing and understanding of spoken language. In particular, broadcast news highlighting is related to speech summarisation, gisting, information retrieval and soundbite detection.

Speech summarisation as a research discipline is relatively new. To date it has attracted less attention than the parallel field of text summarisation, which can be partly attributed to the challenging nature of the task involving the distillation of an audio signal to a textual summary. Some successful speech summarisation systems have been documented. In the early work by Valenza *et al*, features inspired by information retrieval techniques were applied to select words from ASR transcripts to include in a summary [1]. Similar ideas were pursued in [2] and [3], where an automatic summarisation system was proposed, at the core of which was a sentence-by-sentence compression module. Other works on broadcast news data include [4]–[6]. Related tasks such as summarisation of multiparty meetings, voicemails and lectures/talks have also attracted a growing amount of interest [7]–[11]. Broadcast news highlighting is closely related to works on speech gisting and headline generation, where short sentences are generated as part of a speech understanding system [12]–[14]. Inspiration from *text* summarisation efforts such as the 2004 Document Understanding Conference (DUC) headline and very short summary (< 75 bytes) tasks is also called for [15], although care must be taken when porting text-based systems to speech [16]. Works on soundbite extraction, where the user's ability to access spoken data is aided by the provision of informative

This research was supported by EPSRC grant GR/R42405 S3L: statistical summarisation of spoken language.

H. Christensen and Y. Gotoh are with the Department of Computer Science, University of Sheffield, S1 4DP, Sheffield, United Kingdom (email: h.christensen@dcs.shef.ac.uk; y.gotoh@dcs.shef.ac.uk)

S. Renals is with the Centre for Speech Technology Research, University of Edinburgh, Edinburgh, EH8 9LW, United Kingdom (email: s.renals@ed.ac.uk)

and/or indicative soundbites, are also relevant [17], [18].

All of the above systems operating on broadcast news were only applied to *manually* segmented ASR transcripts, at the story level and often at the utterance level. Further the amount of data used was much smaller than that typically used in other spoken language research, and considerably less than that used in recent written document summarisation research. Most broadcast news summarisation studies used around a couple of hundred utterances or even less. To address the lack of a suitable archive, we have annotated the story highlights in a corpus of broadcast news shows containing more than 43 hours of speech and comprising more than 21 000 utterances.

Our previous broadcast news highlighting system was built on manually segmented data [19]. This paper presents a fully automatic system that implements the highlighting stage on *automatically* segmented ASR transcripts. The main focus of this paper is to address the effects of the following on the quality of extracted highlights:

- different degrees of error in the speech-to-text transcripts;
- the erroneous segmentation of ASR transcripts into stories and utterances;
- the robustness of the individual features used for highlighting against the above mentioned irregularities in the system.

The news highlighting system consists of three stages: transcription, segmentation and saliency scoring/highlighting. We have access to the audio stream and corresponding ASR transcripts (and to closed captions for control experiments) for TV broadcast news programmes; previously we have also implemented a topic and utterance segmentation stage, hence we focus on the third stage of the system and on analysing the overall behaviour of the system in different configurations. Paramount to the success of the highlighting stage is robustness of the saliency features to the propagated errors resulting from cascading various component technologies. For example, one can expect that saliency features conveying content information (such as information related to the word frequency statistics) are vulnerable to errors introduced by an ASR system, and that stylistic features (such as the position of an utterance in a story) are affected by segmentation inaccuracies.

The remaining paper is organised as follows: Section II presents the data and annotation schemes. Section III briefly describes the principles of the automatic news programme segmentation, that are based on exponential models and the maximum entropy principle. The extractive highlighting stage, presented in Section IV is a feature based approach using a set of multi-layer perceptron (MLP) classifiers. Section V outlines the evaluation scheme, together with the scheme for annotation of 'gold-standard' extractive highlights. The experimental results and extensive analyses of system outputs are shown in Section VI. Finally, conclusions drawn from the work are presented in Section VII.

II. Data

This paper is centred on the application and usability of statistical models for information extraction in fully automatic systems. The availability of suitable data is therefore a crucial element and we have annotated a corpus of over 850 broadcast news stories with extractive highlights. We chose to concentrate on a single news source, rather than spreading efforts over a number of sources. We annotated a set of 114 ABC news broadcasts from the TDT–2 corpus[1] totalling 43 hours of speech. Each program, spanning 30 minutes as broadcast, was reduced to around 22 minutes once commercial breaks were removed[2], and contained on average 7–8 news stories per broadcast, with a total of 855 stories in the 114 broadcasts. In addition to the acoustic data, both manually generated closed caption (word error rate (WER) 13.5%) and six ASR transcripts, with WERs ranging from 20.5% to 32.0%, were available [21]. In our previous work on news stream segmentation, experiments were carried out using the entire set of six ASR transcripts [16]. It was found that the WER variations over the six transcripts made no significant difference, and hence for this work we have chosen to compare only the closed captions with the most accurate ASR transcription (`cuhtk-s1`) [22], the latter being closer in performance to the most recent state-of-the-art ASR systems.

As part of the TREC/SDR evaluations, individual news stories from all closed caption transcripts were segmented by hand. Further, we manually annotated for utterance boundaries. For control experiments, the above hand segmentations were imposed on the ASR transcripts through alignment with the closed captions. All alignments were made automatically but manually examined to minimise the number of errors.

	#news programs	#hours	#utterances	#words
Train	95	35.8	17 948	235 593
Dev	9	3.4	1679	22 871
Test	10	3.9	1966	24 996
Total	114	43.1	21 593	281 460

TABLE I
STATISTICS FOR THE TRAINING, DEVELOPMENT AND TESTING PART OF THE DATA SET.

Table I shows statistics on the partitions of the data set. Statistical models for story segmentation, utterance segmentation and extraction were derived using the common training set. All experimental results reported in this paper are measured using the testing data. Additionally, in order to train and evaluate the statistical models, all news programs were annotated with a gold-standard, one utterance highlight, containing the utterance considered the single most important in each news story. Further details on the creation of these highlights, and the evaluation scheme in general, are given in Section V.

III. Stream Segmentation

Applications such as highlight extraction, headline generation, news archive browsing, or query-based information retrieval often rely on the availability of structured broadcast news data. The audio news stream needs to be processed in

[1]The TDT–2 corpus [20] has been used in the NIST Topic Detection and Tracking evaluations and in the TREC–8 and TREC–9 spoken document retrieval (SDR) evaluations.

[2]Commercial content was filtered out on the basis of annotations provided with the original data.

order to provide typographic cues (such as punctuation, named entity capitalisation and paragraphs) and to be partitioned into coherent units (such as utterances and stories). This section presents the segmentation stage of the system, where utterance and story boundaries are identified using the maximum entropy framework [23], [24].

A maximum entropy model incorporates any prior set of statistical constraints (or feature functions) f concerning the target distribution, and otherwise assumes a uniform probability distribution. In text segmentation the context X can be assumed unique at each boundary, and the feature functions take the form of binary questions. An example of a feature function may be

$$f(X) = \begin{cases} 1 & \text{if, for the boundary context } X, \text{ a word stem (e.g. '}\textit{today}\text{') appears in the utterance before the boundary;} \\ 0 & \text{otherwise.} \end{cases}$$

The number of such feature functions is large, and a common practise is to precede the model training with a feature selection stage. In [24] a greedy feature selection algorithm was employed in order to choose the features exhibiting the largest gain for the model. For computational reasons the approximate gain was used in our implementation, and we have additionally implemented the fast feature selection algorithm proposed in [25]. This method, 'the Selective Gain Computation Algorithm', further speeds up the feature selection by limiting the number of gain calculations.

The story boundary model treats utterances as units, and the model provides statistics for assigning a probability to each utterance indicating to which degree it is the last utterance before a story boundary. It relies on both lexical and prosodic information, using three distinct types of feature function, of which two types are cue word based and one is derived from pause duration. The architecture of the utterance boundary detector is in principle similar to that of the story boundary model. However it operates on a word level, thus hypothesising each word as a utterance boundary candidate. It is based only on prosodic information (pause) which we investigated in the previous work [26], [27]. It generalises well to the wide spectrum of speaking and language styles found in broadcast news, ranging from read anchor speech to acoustically challenging, spontaneous speech from the field interviews.

For story boundary modelling the feature selection algorithm was used to reduce the large number of cue word features (47 500 in total) to a more manageable size. Preliminary experiments on the closed caption transcripts showed that selecting around 100 features resulted in a reasonable balance between calculation speed and performance. The same number of features was used by [24]. No feature selection was needed for utterance segmentation because there were only a small number of pause features.

IV. Saliency Scoring

The task of the saliency scoring stage is to automatically generate a highlight for a broadcast news program, consisting of a collection of the most important utterances — one for each news story. The problem of selecting a single utterance from a speech transcript is somewhat similar to that of generating a headline. However, while headlines tend to be more compact and can be classified as being either *eye-catchers*, *indicative* or *informative*, no length restriction is set on the highlights in this work. The saliency scoring aims to identify utterances with a high degree of information about their corresponding news story. The approach uses a feature based model that assigns a score to each utterance, indicating how suitable that utterance is for inclusion in a highlight. It is an entirely trainable approach that is suitable when dealing with large amounts of data with varying compositions. Further, it is robust to the origin of utterances being an anchor, a reporter or an interviewee. It was found that 25%, 49% and 26% of utterances were attributed to either 'anchor', 'reporter' and 'others' (e.g., interviewees), respectively. However, this distrubution varies greatly from story to story; it was found that a large number (25%) of stories, typically shorter ones, are delivered by the anchor alone [28].

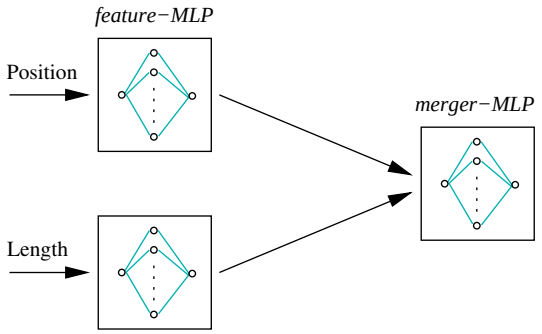

Fig. 1. The two-level architecture of MLP classifiers. All MLPs have 20 hidden units in a single hidden layer.

A. Multi-Layer Perceptron Architecture

The highlighting component was based around a set of multi-layer perceptron (MLP) classifiers [29]. Figure 1 shows the two-level architecture adopted in this study. The first level MLPs (*feature-MLPs*) process individual features derived from each utterance. The second level MLP (*merger-MLP*) combines the outputs of the *feature-MLPs*. The approach is similar to that employed by [30], in which a Naive Bayes classifier was implemented using discretised features. Each *feature-MLP* is trained with a single feature to optimise the utterance selection. The training set consists of a set of features and a gold-standard label ('1' for selected, and '0' for not selected) for each utterance. The outputs of the *feature-MLPs* constitute a vector, which is in turn used as input to the *merger-MLP*. This two-level architecture was chosen primarily because it facilitates the analysis of the contribution from each feature, by sampling the performance of the *feature-MLPs*.

B. Features for Utterance Extraction

In [19], we investigated a large set of candidate features, which can be classified into four categories:

- position of the utterance in the story,

feature	description
POSITION	reciprocal position of the utterance within a story
LENGTH	length of the utterance in number of words
TF.IDF	mean of normalised *tf.idf* terms
COSINE	similarity of *tf.idf* terms between the utterance and the story

TABLE II
THIS TABLE SHOWS A SET OF FOUR FEATURES USED FOR UTTERANCE SELECTION, REPRESENTING INPUTS TO *feature-MLPs*.

- length of the utterance,
- similarity of the utterance to the overall document, and
- distribution of named entities (NEs) within the utterance.

For this work we have settled on a set of four features listed in Table II: POSITION, LENGTH, TF.IDF, and COSINE. The first two features may be classified as structural features, and are concerned with the length and the position of the utterance within a news story. The remaining two features are related to the content of the utterance. The 'TF.IDF' feature is based on traditional information retrieval parameters, comprising the *tf* (*term frequency*) and the *idf* (*inverse document frequency*) statistics. The COSINE feature is the similarity measure of the *tf.idf* vector for the utterance with that for the entire story [31].

V. EVALUATION SCHEMES AND GOLD-STANDARD ANNOTATION

The evaluation of a highlighting system is a non-trivial problem. In the speech summarisation community, evaluation is currently the focus in a number of studies [32]–[34]. The main problem is that the notion of what constitutes a 'good highlight' is very subjective; to achieve a consensus as to which information in a story is useful for skimming and ought to be extracted, a large number of human highlights needs to be assembled. The optimal situation would be access to such a pool of gold-standard utterances for each new story. This would enable scoring methods like e.g. the Relative Utility (RU) method of Dagromir Radev [35], where all judges give a utility score to each utterance in the document, in effect ranking them, and where the overall score for a given extractive set of utterances is related to the inter-judge agreement. Another often used evaluation method enabling the scoring against multiple judges is the ROUGE method [36], which is widely used in, mostly non-extractive, summarisation research like the DUC efforts [15]. ROUGE is based on counting overlapping N-grams of text between judge and machine candidates, and hence is less suitable for extractive methods like the one used in this paper. Overall, given the scale of the task of establishing multiple annotations, let alone complete rankings of all utterances in a story as required for the RU method, a single annotator was chosen after a verification stage as described below.

A. Gold-Standard Annotations

In order to evaluate and train the statistical model on which the saliency scoring stage is founded, a gold-standard, one-utterance extractive highlight for each story is necessary.

	$\hat{\kappa}$
Short stories	0.82
Long stories	0.34
All stories	0.56

TABLE III
THIS TABLE SHOWS THE ESTIMATED $\hat{\kappa}$ VALUES QUANTIFYING THE DEGREE OF AGREEMENT BETWEEN SIX PEOPLE FOR 44 RANDOMLY SELECTED STORIES. $\hat{\kappa} = 0.82$ (SHORT STORIES) AND 0.34 (LONG STORIES) INDICATE THE VERY STRONG AND MODERATE AGREEMENTS, RESPECTIVELY.

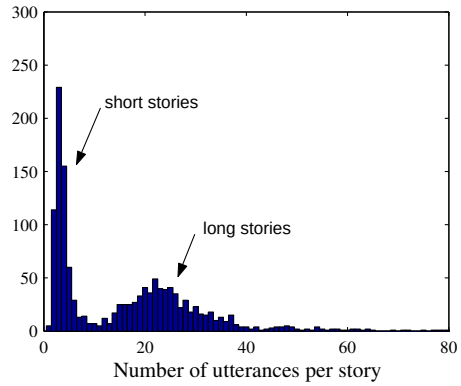

Fig. 2. The bar graph shows that the story length over the entire corpus (855 stories) roughly adhered to a bimodal distribution. With the median length of 16 utterances, they may be split into short and long stories.

To this end, a human selected one utterance from each of the 855 news stories in the closed caption transcripts. They were then examined for their quality and consistency in the following manner: five additional people individually selected an utterance from each of the 44 news stories (a small subset randomly chosen and spread in time throughout the corpus). To assess the level of agreement between the six people the κ (kappa) statistic was used [37]. κ is a measure of agreement between subjects, taking into account the agreement one would expect to see arising from pure chance. It can be estimated by

$$\hat{\kappa} = \frac{\hat{P}(A) - \hat{P}(E)}{1 - \hat{P}(E)} \tag{1}$$

where $\hat{P}(A)$ is the estimate of the proportion of inter-subject agreements and $\hat{P}(E)$ is the estimate of the expected proportion of chance agreements. $\hat{\kappa}$ is '1' when all subjects are in perfect agreement, and '0' when there is only a chance agreement. Table III shows the estimated $\hat{\kappa}$ values quantifying the degree of agreement between six people. The overall average was 0.56, indicating a good agreement — sufficiently good for a gold-standard. It was therefore chosen to use one annotator for the full corpus.

Estimation of $\hat{\kappa}$ according to Equation (1) assumes a *constant* number of categories that each person can choose from. However in this case, the number of categories (i.e., the number of utterances in a given story) varied from story to story. As the graph in the Figure 2 indicates, the story length over the entire corpus roughly adhered to a bimodal distribution. There were many very short stories (typically 2–5 utterances)

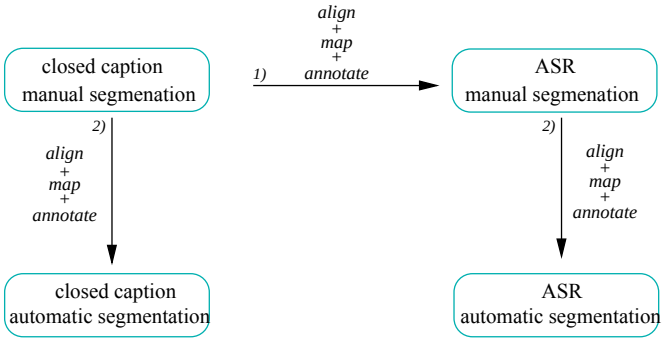

Fig. 3. Illustration of the approach for obtaining gold-standard annotations for all experimental conditions.

and many longer stories (20–30 utterances), the latter leaving room for much more disagreement between people. A separate $\hat{\kappa}$ measure was therefore calculated for stories below and above the median story length of 16 utterances; the estimated $\hat{\kappa}$ values were 0.82 for short stories, and 0.34 for long stories indicating very strong and moderate agreements respectively.

Not surprisingly, people had a very high level of agreement for short stories, but less so for longer stories, indicating the difficulty of the task. They commented that to select a single utterance to represent an entire news story was at times very difficult. The observed disagreements can partly be attributed to the difficulty in defining (and interpreting) what is meant by a suitable one-utterance highlight of a news story. Below is an example, containing the two lead utterances of a news story, where both utterances can reasonably be considered as a good candidate:

> ***1:*** *in texas karla faye tucker is running out of time*
> ***2:*** *she is the convicted killer at the center of a debate about repentance and the death penalty*

The first utterance contains the location (*texas*) and the person name (*karla faye tucker*), whereas the second utterance gives more specific information about the content of this news story. The latter would provide valuable information to distinguish this particular story from similar stories containing further or previous developments in the case, since all the stories are likely to mention the same location and person.

Obtaining gold-standard for all experimental conditions. The experiments in this study were conducted under various scenarios using different combinations of manual and automatic segmentations and ASR qualities, hence requiring 'gold-standard' annotations for all scenarios. Using the high-quality, gold-standard highlights described above for the closed caption as a starting point, annotations for the remaining scenarios were obtained by way of a three-stage procedure:

1) align the transcripts from different scenarios on a word basis;
2) map annotations from one scenario to the other;
3) annotate, as the new gold-standard, the utterance having the most overlap (in terms of words) with the original highlight.

As Figure 3 shows, this 'align+map+annotate' procedure was initially applied to map annotations from the closed-caption with manual segmentations to the ASR transcripts also with manual segmentations. Then it was applied to map from these manually segmented set of transcripts (closed-caption and ASR respectively) to their automatically segmented counterpart transcripts.

B. Evaluation

The use of both intrinsic and extrinsic evaluation measures for this kind of system is warranted [38]. The automation and speediness of intrinsic evaluation measures make them suitable for testing and developing many system configurations on large amounts of data, whereas extrinsic techniques like user tests are resource heavy, but can provide a more refined picture of the performance. In order to establish the usefulness of having access to a one-utterance highlight for each broadcast news transcript, and to quantify to which degree the fully automatic system is able to return informative highlights, two rounds of user tests have been conducted. The setup for the extrinsic evaluation is described further in section VI. As for the intrinsic evaluations, some of the factors complicating the evaluation process have been reduced because the highlights are extractive and because a good inter-human agreement has been measured on the gold-standard against which the quality of highlights are measured. Thus the use of information extraction related measures, such as precision and recall, is merited. In the experiments described in the next section, receiver operating characteristic (ROC) analysis is also used extensively. An ROC curve depicts the relation between the true positive and the false positive rates for every possible threshold.

However, for automatic segmentation scenarios used in this work, caution is required because the gold-standard is not perfect, but instead obtained using the mapping process described in Section V-A. In evaluation terms, this means that only if the highlighting system picks the exact gold-standard utterance will it count as 'correct' (i.e., increase the true positive score), whereas if the utterance before or after is picked (which is also likely to contain a (shorter) snippet of the original gold-standard highlight), this is not credited; rather it will contribute to the false positive score.

Using the above evaluation scheme, ASR transcription errors will not affect the ROC analysis directly, however it might still have an adverse effect by causing failures in the story and utterance segmentation components.

VI. Experiments

The research issues, outlined in Section I, concern the effect on the quality of extractive one-utterance highlights, when applied to various combinations of transcription, utterance and story segmentation quality. Table IV provides an overview of the factors involved. Our experiments used manual and automatic utterance segmentation (labelled U_m and U_a) and manual and two forms of automatic story segmentation (labelled S_m, S_{a1}, S_{a2}). S_{a1} applied cue word features only, whereas S_{a2} also incorporated pause information.

We begin this section with a set of extrinsic evaluations that demonstrate the usefulness of having access to one-utterance highlights when skimming broadcast news transcripts.

Factor	Variations
ASR quality	closed caption (CC), ASR (ASR)
Utterance segmentation	manual (U_m), automatic (U_a)
Story segmentation	manual (S_m), automatic (S_{a1}, S_{a2})

TABLE IV

THIS TABLE SHOWS THE FACTORS INVOLVED IN THE EXPERIMENTS. TWO AUTOMATIC STORY SEGMENTATIONS WERE TESTED: ONE USING CUE WORDS ONLY (S_{a1}) AND ONE WITH CUE WORDS PLUS PAUSE FEATURES (S_{a2}).

A. Usefulness of highlights

The user test is task-based and involves the subject having to skim through a list of news story transcripts to look for an answer to a question. News stories are individually numbered, and the subject is asked to return the story number believed to contain the answer to the question. As well as monitoring the correctness of the answer, the response time is measured. Assessing usefulness of summaries through timed tasks is a method often employed in summarisation research (see, e.g., [36], [39]).

A set of 12 target stories with either manual or automatic topic and utterance segmentations were selected at random from those containing more than 10 utterances, and for each target story a question was formulated. The questions were constructed without knowledge of which utterance had been extracted as the gold-standard or automatic highlight. They would query a certain fact specific to that particular news story, i.e. for a story unfolding over several days, the question would be related to the new development reported. Examples are:

- How are the IRS treating divorced spouses who's ex-husbands have large tax debts?
- What problems do the UN weapon inspectors have with getting access to sites in Iraq?
- Which tactics have policed employed in the hunt for the suspected cop killers?

The degree of difficulty faced by the subjects in this type of task is highly dependent on the confusability of the stories, i.e. the content overlap between individual stories and between question and target/distractor stories is relevant. We therefore employed a vector space framework and measured the cosine distance between vectors consisting of tf.idf weights [31]. The results showed that the subjects faced little content overlap with a cosine dist, $cosDist$ for inter story overlap being $cosDist_{interStory} < 0.1217$. As expected, the questions' overlap with the target story is larger than the overlap with the distractor stories, $cosDist_{q \leftrightarrow target} = 0.0691$ and $cosDist_{q \leftrightarrow distractor} = 0.0016$.

The tests were conducted with the aid of a computer, and the software interface displayed the questions to the subject one at a time. For each question, the transcripts of 17 news stories were presented on the screen, each clearly marked with a story number; one of the stories was the target story and the remaining 16 distractor stories were chosen at random. Stories were presented in one of six different styles: 1) the complete transcripts, displayed one after the other with clearly marked story numbers ('full stories'); 2) like the first style but with the highlight utterance in each story emphasised in red font ('full stories + HL'); 3) again like the first style, but with keywords in the form of named entities[3] emphasised in red font ('full stories + KWs') 4) displaying only the highlight utterance for each story ('HL only'), 5) displaying only the keywords for each story ('KWs only') and 6) presenting an utterance picked at random from each story. An example of the user interface is given in figure 8 in Appendix.

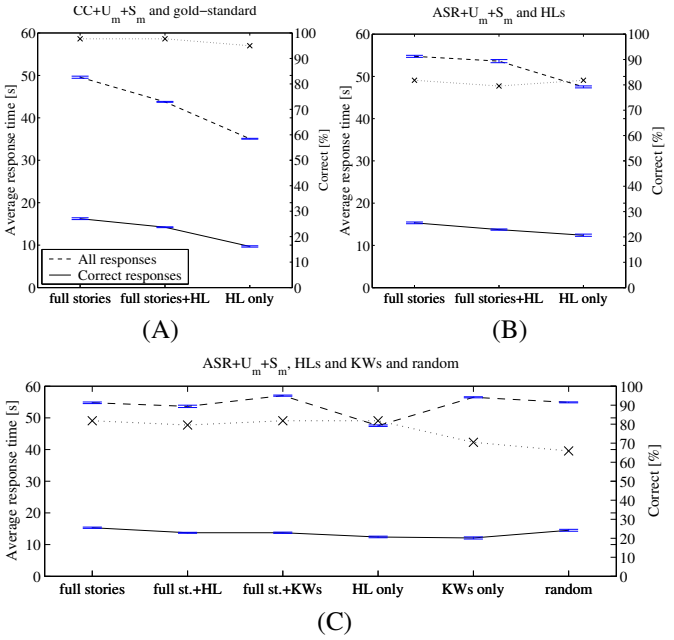

Fig. 4. (A) and (B): Average response time for the three different presentation styles involving highlights in the user test (against left-hand axis) and % correct (against right-hand axis). Errorbars represent the standard error of the mean. They are very small and the horizontal bars have been amplified to make them more detectable. Dashed line and solid line represent the response time for all and correct responses, respectively. (C): Average response time for the six different presentation styles involving highlights, keywords and the random utterance extraction system, axes as in (A) and (B).

The presentation style alternated with question and we controlled for question/presentation style pairings by devising three different versions of the user test, which subjects were assigned to at random. A total of 11 subjects participated in the test, all of whom were experienced in skimming text and looking for information on-screen. The subjects were given full training in how to use the software interface by conducting a dummy experiment, which contained stories not present in the formal test. Thus they were familiar with the task and confident in using the software. The subjects were informed that their response time was being measured. To encourage the right behaviour of trying to skim the text rather than to read the full stories, the subjects were competing for a prize for the best conducted experiment (i.e. fastest and most correct entry). The full instruction, given to the subjects, is shown in Appendix.

On Figure 4, plot A shows the results for the reference system (manual story and utterance segmentation of closed caption transcripts and with gold-standard highlights — $CC+$

[3]Named entities were extracted automatically following Gotoh and Renals [40].

$U_m + S_m$). Panel A and B present results for the three presentation styles 'full stories', 'full stories + HL' and 'HL only'. The average response time for the three different presentation styles for all responses and for correct responses only are given against the left-hand y-axis, and the corresponding % correct answers are plotted against the right-hand y-axis. Having the gold-standard utterance highlighted in the manually segmented stories sped up the task significantly, and the fastest average response times were seen for the the 'HL only' case. Similarly, plot B shows results for the fully automatic system (automatic story and utterance segmentation of ASR transcripts — $ASR + U_a + S_{a2}$). The outcome was similar to that for the reference system, although the improvement resulting from highlights was not as large as the reference system.

Comparing individual response times, not surprisingly the number of incorrect answers increased from roughly 2% for the all manual, reference system ($CC + U_m + S_m$) to 20% for the fully automatic system ($ASR + U_a + S_{a2}$). Most subjects made no incorrect answer when working with the all manual system, and all subjects made at least one incorrect answer for the fully automatic system. However the most important message is that there is essentially no decline in the number of correct answers by using extracted single utterance highlights instead of full transcripts.

On Figure 4, plot C shows the results from the full set of presentation styles which enables the comparison of the highlighting system to various baseline systems. The response times from plot B are repeated and results from the presentation styles involving keywords and the 'random' systems are added. The response times for 'full stories+KWs' is comparable to those of the other full story presentation styles, and as with the highlighting case, the 'KWs only' system is quicker to use than the full stories systems. The correct response times of 'KWs only' are at level with the 'HL only' systems, however, the users' ability to make a correct judgement is significantly reduced when only keywords are provided; highlighting keywords is a strategy used by many online search engines and audio browsers, e.g. [41], but for this task, highlighting a whole sentence provides the user with a better skimming facility. Finally, comparing with the 'random' presentation style shows that the automatic highlighting system performs significantly better, both in terms of speediness and correctness, than a system highlighting a random utterance.

The user tests have demonstrated that the one-utterance strategy is useful for the broadcast news skimming scenario, and that the implemented fully automatic system is capable of supplying the user with transcripts of a good quality. With this established, the remainder of this section concentrates on the intrinsic evaluations carried out to analyse the effect of the different cascading systems.

B. Effect of Word Recognition Errors

The effect of varying WERs was explored on a parallel set of transcripts: the closed captions (CC) with the WER 13.5% and the ASR transcripts (ASR) with the WER 20.5%. Figure 5 shows their ROC curves under three characteristic segmentation scenarios.

On all three plots, the ROC curves for CC and ASR were very similar. This is also reflected in the F-measure numbers that differ only a maximum of 0.04. It is important to bear in mind that this proximity only implies the saliency scoring stage based on this particular ASR transcript was as good as the closed caption at predicting the gold-standard highlight. However, this type of evaluation does not assess how suitable the selected one-utterance highlight is at condensing and expressing the given news story. The chosen gold-standard utterance is the one overlapping the most with the CC gold-standard, and it may be considered a reasonable representation of that news story as it has the structural, and partly lexical similarity with the original gold-standard.

However, as we observed in [19], by using humans to judge the quality of the resulting highlights, the suitability of an extract from an ASR transcript is lower than what may be suggested by the ROC analysis. In [19], although the ROC curves were similar, the human judges expressed a clear preference for the CC highlights over those from ASR. This was probably due to the different nature of the transcription errors in CC and ASR. Errors in the manual transcripts were made by humans and often occurred because the stenocaptioner had misheard or possibly was unable to keep up with the broadcast. Occasionally a complete phrase was lost, but the grammatical structure and contents were intact for the majority of utterances. On the other hand, ASR errors were spread out; it is easy to imagine that even a single word substitution, with a small impact on the WER (e.g., "is" to "isn't"), can change the meaning of an utterance completely. However, as indicated by the user test discussed in Section VI-A, although the highlight from automatic transcripts may appear less grammatical and readable, when skimming a text, they are still highly useful.

C. Effect of Story and Utterance Segmentation Errors

Figure 6 shows the ROC curves for extractive highlights under various combinations of manual and automatic segmentations, ranging from manual segmentations for both story and utterance boundaries ($U_m + S_m$) through to a fully automatic system ($U_a + S_{a2}$).

Table V presents the precision and recall scores, and their harmonic mean (F-measure), for a few key systems[4].

For both the closed captions (CC) and the ASR transcripts (ASR), the ROC curves were clearly separated into two groups. The best performing curves were for those combinations using manual story segmentation; $U_m + S_m$ and $U_a + S_m$ combinations. F-measures were ranging from 0.35 to 0.61. The rest of the combinations, arising from automatic story segmentation, performed significantly worse with F-measures between 0.23 and 0.31. The type of automatic story segmentation — cue words only (S_{a1}) or cue words plus pause (S_{a2}) — made very little difference in highlighting performance. Interestingly, automatic utterance segmentation did not affect the highlight extraction adversely.

[4]The operating point (OP) for which the scores are calculated is determined by inspecting the relevant ROC curve; the OP is chosen as the system with the maximum area under the curve.

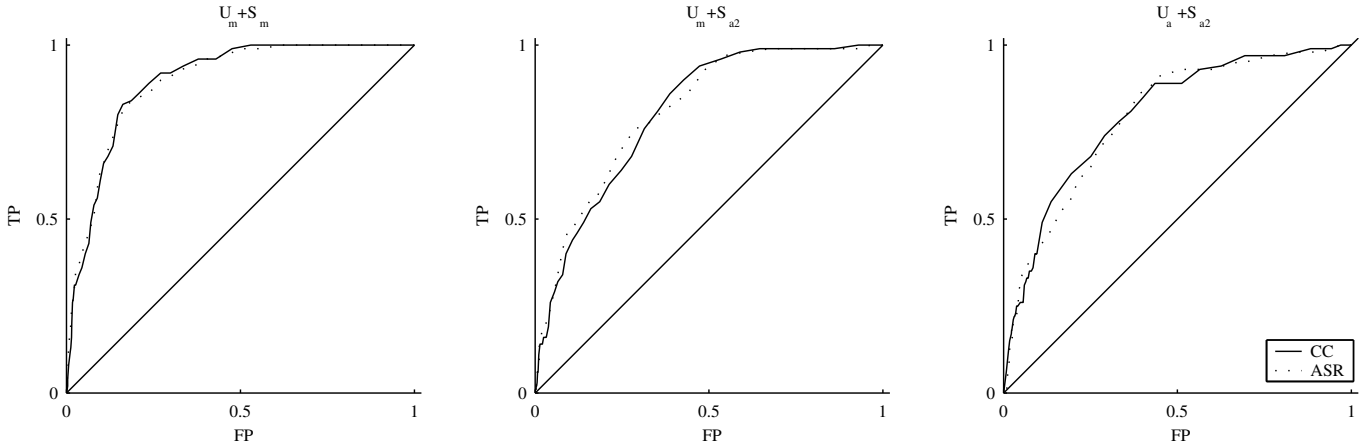

Fig. 5. ROC curves presenting the highlighting performance for the closed captions (CC) with WER 13.5% and the ASR transcripts (ASR) with WER 20.5% under three characteristic segmentation scenarios, $U_m + S_m$, $U_m + S_{a2}$ and $U_a + S_{a2}$.

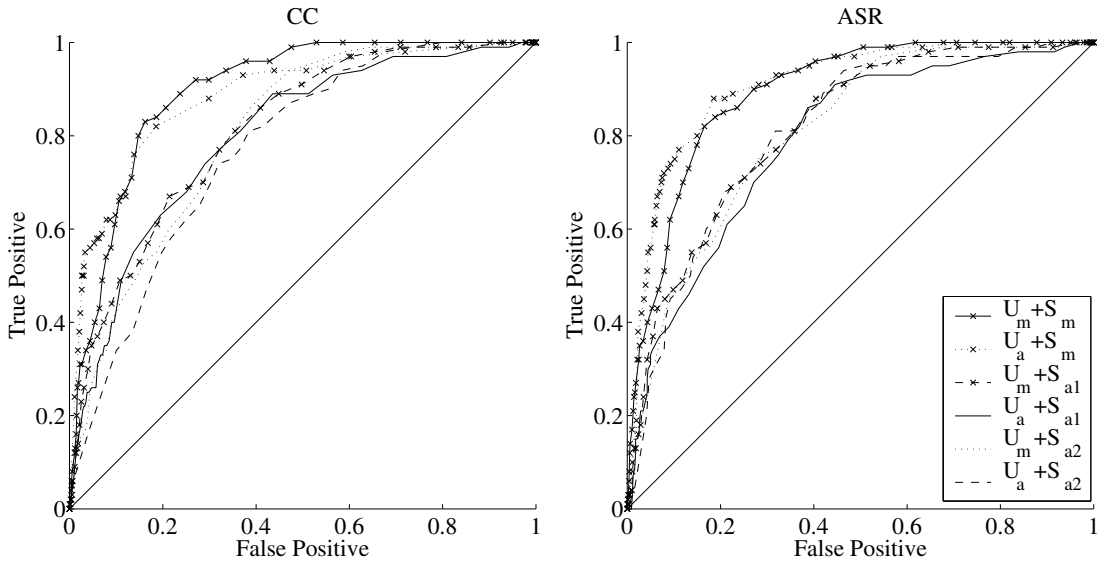

Fig. 6. ROC curves presenting the highlighting performance for the closed caption (CC) and the ASR transcript (ASR) using the different combinations of manual and automatic segmentations.

CC			
System	Recall	Precision	F-measure
$U_m + S_m$	0.83	0.30	0.44
$U_a + S_m$	0.82	0.22	0.35
$U_m + S_{a2}$	0.70	0.17	0.27
$U_a + S_{a2}$	0.74	0.14	0.23
ASR			
System	Recall	Precision	F-measure
$U_m + S_m$	0.82	0.30	0.44
$U_a + S_m$	0.80	0.49	0.61
$U_m + S_{a2}$	0.74	0.18	0.29
$U_a + S_{a2}$	0.70	0.20	0.31

TABLE V

THIS TABLE LISTS THE RECALL, PRECISION AND F-MEASURE SCORES USING THE DIFFERENT COMBINATIONS OF MANUAL AND AUTOMATIC SEGMENTATIONS; THE SCORES ARE CALCULATED AT THE OPERATING POINTS ON THE ROC CURVES CLOSEST TO THE TOP RIGHT-HAND CORNER GIVING THE MAXIMUM AREA UNDER THE CURVE.

What might explain this drastic degradation in performance when the manual story segmentation was replaced with the automatic approach ($S_m \rightarrow S_a$)? Either of the four stages could hold the answer: it could be that the automatic story segmentation performed very badly, it could be that the features in the saliency scoring were more vulnerable to a story segmentation error, or it could be caused by interactions with the utterance segmentation and/or transcription stages. Table VI presents an overview of the performances in F-measures for the transcription and two following segmentation, i.e., after the utterance segmentation only, or after the utterance segmentation followed by the story segmentation[5].

[5]Both the utterance and story segmentation stages were tuned on the development set, so as to output approximately the right number of boundaries. Further, the manual story segmentations were enforced to the nearest automatic utterance boundary to provide segmentations for the $U_a + S_{a2}$ scenario; the low resulting score is due to the target boundaries being kept the same.

Transcription	Utterance seg.		Story seg.	
CC (WER: 13.5%)	manual	1	manual	1
			auto	0.46
	auto	0.07	manual	0.14
			auto	0.14
ASR (WER: 20.5%)	manual	1	manual	1
			auto	0.53
	auto	0.05	manual	0.14
			auto	0.12

TABLE VI

ANALYSIS OF PERFORMANCE OF CASCADING STAGES: TRANSCRIPTION QUALITY, UTTERANCE SEGMENTATION AND STORY SEGMENTATION. NUMBERS ARE F-MEASURES WHEN SCORED AGAINST THE CORRESPONDING MANUAL ANNOTATIONS. TO MAKE THE TABLE COMPLETE THE SCORES FOR THE MANUAL ANNOTATION STAGES (1) ARE ALSO GIVEN.

System	Feature	Recall	Precision	F-measure
$U_m + S_m$	POSITION	0.76	0.21	0.33
	LENGTH	0.73	0.30	0.40
	TF.IDF	0.70	0.30	0.26
	COSINE	0.66	0.30	0.40
	comp	0.82	0.30	0.44
System	Feature	Recall	Precision	F-measure
$U_m + S_{a2}$	POSITION	0.64	0.13	0.22
	LENGTH	0.73	0.17	0.25
	TF.IDF	0.69	0.23	0.23
	COSINE	0.63	0.17	0.25
	comp	0.74	0.18	0.29
System	Feature	Recall	Precision	F-measure
$U_a + S_{a2}$	POSITION	0.66	0.13	0.21
	LENGTH	0.70	0.16	0.26
	TF.IDF	0.66	0.16	0.24
	COSINE	0.64	0.17	0.27
	comp	0.70	0.20	0.31

TABLE VII

THIS TABLE LISTS THE RECALL, PRECISION AND F-MEASURE SCORES FOR THREE CHARACTERISTIC SCENARIOS AND ON ASR TRANSCRIPTS.

It shows the same, very little degradation between transcription quality changes ($CC \rightarrow ASR$; comparing the top half of the table to the bottom half), as is still observed when probing after the highlighting stage (see Section VI-B). It also shows that in terms of F-measure, the degradation in performance after the utterance segmentation is actually worse than that after the story segmentation. However, the reason that this does not translate into abysmal performance at the highlighting stage (as both extrinsic and intrinsic results have shown), or indeed changes greatly in terms of how the story segmentation performs, is the pause cue itself. Pause duration is particular good at indicating a story boundary, and it is also a direct consequence of the prosody of the speaker and hence makes for good unit boundaries in terms of being intelligible, conceptually sound and conveying coherent information. The culprit must be found in the ultimate stage of the system, and Section VI-D is dedicated to analysing in more detail, the contribution of the individual saliency features.

D. Effect of Individual Features Contribution

Figure 7 illustrates the contribution of the individual features for extractive one-utterance highlighting. It represents three different segmentation scenarios, $U_m + S_m$, $U_m + S_{a2}$ and $U_a + S_{a2}$ — all with the ASR transcripts. Table VII presents the precision, recall and F-measure scores for the similar scenarios for each of the individual features as well as their MLP-combination.

As expected, the highlighting system combining the features (the line labelled with '*MLP comb.*') outperformed those employing just a single feature. Interestingly, the amount of performance gain achieved, illustrated in terms of the gap between the 'MLP comb.' curve and the one with the best single feature varied significantly depending on the type of segmentation scheme employed. In the $U_m + S_m$ scenario (i.e., manual segmentations for both utterances and stories — left-hand side panel in Figure 7) all features were contributing complementary information, resulting in a much better ROC curve when combined. However, this was not the case, when the story segmentation was automatic, as seen in the remaining two panels ($U_m + S_{a2}$ and $U_a + S_{a2}$).

This discrepancy might be largely attributed to the degradation in performance of the POSITION feature. When the story boundaries become erroneous, structural information, such as the position of an utterance within a news story, invariably become less reliable. The careful examination of the three panels in Figure 7 clearly indicates that, when using the manual segmentation ($U_m + S_m$), the POSITION feature was almost as good as the full combination of features, while its contribution had declined significantly for automatic story segmentations ($U_m + S_{a2}$ and $U_a + S_{a2}$). Unreliable segmentations for story and utterance caused the POSITION feature to lose its ability to boost the performance.

This is in contrast with the LENGTH feature. It was a more robust feature, as indicated in Figure 7; the performance of the single feature highlighting fell only slightly, when the segmentation was fully automatic. This is an interesting finding, given that the utterance boundary segmentation was relatively rudimentary, making use only of the pause information. The manual segmentation was based on a grammatical principle, whereas the automatic segmentation was purely guided by the speaker's pattern of pauses. When a speaker was reading the news (e.g., an anchor), pauses were often inserted asynchronously to full stops; they were occasionally used for emphasising certain parts of the story. It may be concluded that the pause-driven automatic segmentation scheme did not result in a reduction in highlighting performance.

Returning to the analysis of the errors cascading through in the case of automatic story segmentation, then the experiments reported in this section have clarified that the POSITION feature is largely responsible for the degradation in performance of the $S_m \rightarrow S_a$ situation.

As a final note, it is valuable to compare the highlighting system against a baseline system. For printed news summarisation, it is widely accepted that a simple, and reasonably good, summarising approach is to let the summary be comprised of the first sentence of a news story. This is also a natural choice as a baseline for experiments with a news broadcast highlighter [19]. The POSITION feature essentially provided

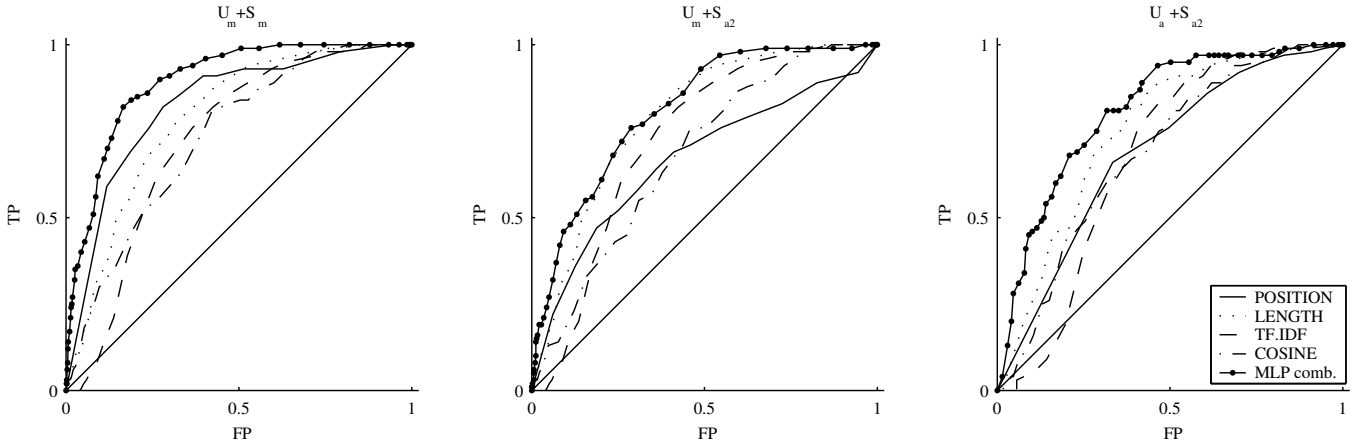

Fig. 7. ROC curves presenting the overall highlighting performance (the line labelled with '*MLP comb.*') and the performances of four individual features (the rest).

this baseline. According to the definition (i.e., the reciprocal position of the utterance within a story), this feature will take its highest value (1) for the first utterance. When using fully automatic transcription and segmentation the highlighting system performs better than this baseline system.

VII. Conclusions

We have presented a fully automatic approach to extractive highlighting of broadcast news programs. The system transcribes the news stream using a speech recogniser, and subsequently segments the data into utterances and stories. Finally it extracts, from each story, one utterance that is the most suitable for a news highlight. Statistical approaches were used for implementation of the system — a maximum entropy framework for the segmentation stage and a multilayer perceptron model for the highlighter.

In this paper, we focused on an essential issue specific to *speech* highlighting — namely, the ability of the highlighting system to operate on data corrupted with segmentation and transcription errors. Experimental results indicated that mistakes in the story boundary detection had a strong impact on the quality of highlights, whereas erroneous utterance boundaries caused only minor problems. Further, the difference in transcription quality between closed captions and ASR outputs did not affect the highlighting performance greatly.

We investigated the robustness of the highlight features to cascading errors in the system. It was found that the content based features were largely unaffected by propagated errors, but that the structural features could not escape from the adverse effects; in particular the position of the utterance in a news story lost its superiority when the story boundaries became less reliable. Finally, user tests demonstrated the overall usefulness of having highlights available when skimming news stories.

The overall conclusion is that a high degree of accuracy can be achieved with a fully automatic highlighting framework, but that the improvement of the story segmentation component is key to boost the quality of broadcast news highlights. Future work should explore other machine learning techniques suitable for implementing the saliency scoring stage to achieve a highlight. Also, the advantages of loosening the one-utterance restriction should be investigated; it is possible that highlights comprised of a couple of sentences would further increase the informativeness.

References

[1] R. Valenza, T. Robinson, M. Hickey, and R. Tucker, "Summarisation of spoken audio through information extraction," in *Proceedings of the ESCA Workshop: Accessing Information in Spoken Audio*, Cambridge, 1999, pp. 111–116, (http://svr-www.eng.cam.ac.uk/~ajr/esca99/).

[2] C. Hori, S. Furui, R. Malkin, H. Yu, and A. Waibel, "A statistical approach for automatic speech summarization," *EURASIP Journal on Applied Signal Processing*, vol. 2, pp. 128–139, 2003.

[3] ——, "Automatic speech summarization applied to english broadcast news speech," in *Proceedings of ICASSP 2002*, 2002.

[4] B. Kolluru, H. Christensen, and Y. Gotoh, "Multi-stage compaction approach to broadcast news summarisation," in *Proceedings of Eurospeech 2005*, Lisbon, 2005, pp. 69–72.

[5] S. Maskey and J. Hirschberg, "Comparing lexial, acoustic/prosodic, discourse and structural features for speech summarization," in *Proceedings of Eurospeech 2005*, Lisbon, 2005, pp. 621–624.

[6] K. McKeown, J. Hirschberg, M. Galley, and S. Maskey, "From text summarization to speech summarization," in *Proceedings of ICASSP 2005*, vol. V, Philadelphia, 2005, pp. 997–1000.

[7] K. Zechner, "Automatic summarization of open domain multi-party dialogues in diverse genres," *Computational Linguistics*, vol. 28, no. 4, 2002.

[8] G. Murray, S. Renals, and J. Carletta, "Extractive summarization of meeting recordings," in *Proceedings of Eurospeech 2005*, Lisbon, 2005, pp. 593–596.

[9] K. Koumpis and S. Renals, "Content-based access to spoken audio," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 61–69, 2005.

[10] M. Galley, "Automatic summarization of conversational multi-party speech," in *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06). AAAI/SIGART Doctoral Consortium*, Boston, USA, 2006.

[11] P. Chatain, E. Whittaker, J. Mrozinski, and S. Furui, "Perplexity based linguistic model adaptation for speech summarisation," in *Proceedings of Interspeech 2006*, Pittsburg, USA, 2006.

[12] W. P. Doran, N. Stokes, E. Newman, J. Dunnion, and J. Carthy, "A hybrid statistical/linguistic approach to news story gisting," in *Proceedings of the 27th ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004.

[13] R. Wang, N. Stokes, W. Doran, E. Newman, J. Carthy, and J. Dunnion, "Comparing topiary-style approaches to headline generation," in *Proceedings of the 27th European Conference on Information Retrieval (ECIR-05)*, Santiago de Compstela, Spain, 2005.

[14] B. Dorr, D. Zajic, and R. Schwartz, "Hedge trimmer: A parse-and-trim approach to headline generation," in *Proceedings of Workshop on Automatic Summarization*, May 2003.
[15] http://duc.nist.gov/.
[16] H. Christensen, B. Kolluru, Y. Gotoh, and S. Renals, "From text summarisation to style-specific summarisation for broadcast news," in *Advances in Information Retrieval*, S. McDonald and J. Tait, Eds. Springer-Verlag, 2004, pp. 223–237.
[17] S. Maskey and J. Hirschberg, "Soundbite detection in broadcast news domain," in *Proceedings of Interspeech 2006*, Pittsburg, USA, 2006.
[18] S. Whittaker, J. Hirschberg, B. Amento, L. Stark, M. Bacchiani, P. Isenhour, L. Stead, G. Zamchick, and A. Rosenberg, "Scanmail: a voicemail interface that makes speech browsable, readable and searchable," in *Proceedings of CHI2002 Conference on Human Computer Interaction*, New York, USA, 2002.
[19] H. Christensen, Y. Gotoh, B. Kolluru, and S. Renals, "Are extractive text summarisation techniques portable to broadcast news?" in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2003)*, U.S. Virgin Islands, 2003.
[20] C. Cieri, D. Graff, M. Liberman, N. Martey, and S. Strassel, "The TDT-2 text and speech corpus," in *Proceedings of DARPA Broadcast News Workshop*, Herndon, VA, 1999, pp. 57–60.
[21] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proceedings of the 6th Conference on Computer Assisted Information Retrieval (RIAO 2000)*, Paris, 2000.
[22] P. Woodland, J. Odell, T. Hain, G. Moore, T. Niesler, A. Tuerk, and E. Whittaker, "Improvements in accuracy and speed in the HTK broadcast news transcription system," in *Proceedings of Eurospeech-99*, 1999.
[23] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.
[24] D. Beeferman, A. Berger, and J. Lafferty, "Statistical models for text segmentation," *Machine Learning*, vol. 34, pp. 177–210, 1999.
[25] Y. Zhou, F. Weng, L. Wu, and H. Schmidt, "A fast algorithm for feature selection in conditional maximum entropy modeling," in *Proceedings of the EMNLP 2003*, Sapporo, 2003, pp. 153–159.
[26] H. Christensen, Y. Gotoh, and S. Renals, "Punctuation annotation using statistical prosody models," in *Proceedings of the ISCA Workshop on Prosody in Speech Recognition and Understanding*, Red Bank NJ, 2001, pp. 35–40.
[27] H. Christensen, B. Kolluru, Y. Gotoh, and S. Renals, "Maximum entropy segmentation of broadcast news," in *Proceedings of ICASSP 2005*, Philadelphia, 2005.
[28] B. Kolluru, "Broadcast news processing: Structural classification, summarisation and evaluation," Ph.D. dissertation, University of Sheffield, Sheffield, 2006.
[29] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press, 1995.
[30] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer," in *SIGIR-95: Workshop on Cross-Linguistic Information Retrieval*, Seattle, 1995, pp. 68–73.
[31] C. D. Manning and H. Schütze, *Foundation of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999.
[32] C.-Y. Lin and E. Hovy, "The potential and limitations of automatic sentence extraction for summarization," in *Proceedings of the HLT-NAACL 2003 Workshop on Automatic Summarization*, Edmonton, 2003.
[33] A. Nenkova and R. Passonneau, "Evaluating content selection in summarization: The pyramid method," in *Proceedings of the HLT-NAACL 2004*, Boston, 2004.
[34] A. Nenkova, "Summarization evaluation for text and speech: Issues and approaches," in *Proceedings of Interspeech 2006*, Pittsburg, USA, 2006.
[35] D. R. Radev and D. Tam, "Summarization evaluation via relative utility," in *Proceedings of the 12th International Conference on Information Knowledge Management (CIKM 2003)*, New Orleans, 2003.
[36] B. J. Dorr, C. Monz, S. President, R. Schwartz, and D. Zajic, "A methodology for extrinsic evaluation of text summarization: does ROUGE correlate?" in *Proceedings of the ACL 2005*, Ann Arbor, 2005, pp. 1–8.
[37] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, pp. 37–46, 1968.
[38] K. Spärck Jones, "Automatic language and information processing: Rethinking evaluation," *Natural Language Engineering*, vol. 7, no. 1, pp. 29–46, 2001.
[39] K. McKeown, R. J. Passonneau, and D. K. Elson, "Do summaries help? a task-based evaluation of multi-document summarization," in *Proceedings of SIGIR 2005*, Salvador, Brazil, 2005.

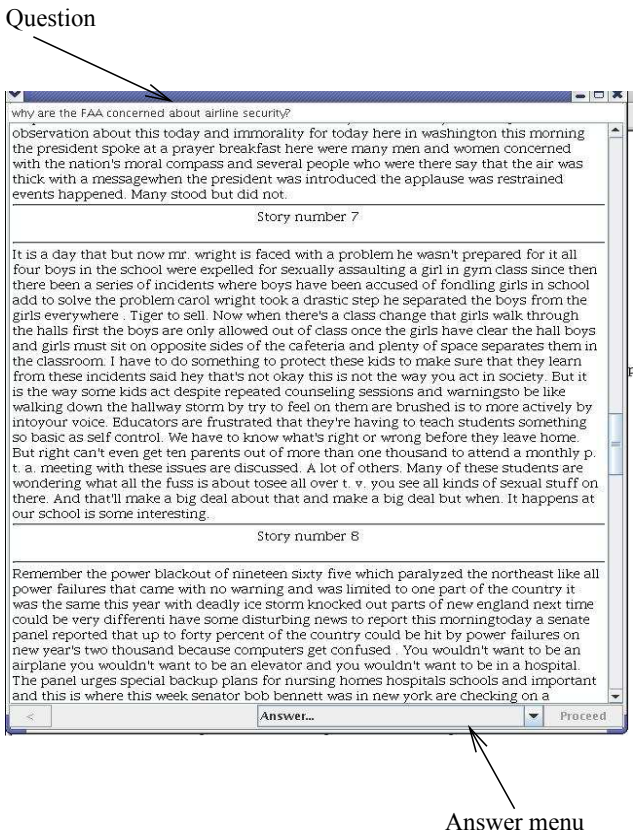

Fig. 8. A screen shot for the user test software, showing full stories. The question is displayed at the top, and the user can submit an answer from the menu at the bottom of the screen. The response time is registered when the 'Proceed' button is clicked.

[40] Y. Gotoh and S. Renals, "Statistical annotation of named entities in spoken audio," in *Proceedings of the ESCA Workshop: Accessing Information in Spoken Audio*, Cambridge, 1999, pp. 43–48, (http://svr-www.eng.cam.ac.uk/~ajr/esca99/).
[41] http://http://web.sls.csail.mit.edu/lectures/.

APPENDIX

Figure 8 shows an interface that subjects uses for extrinsic evaluation of skimming broadcast news stories. The full instruction, given to subjects, reads as:

> *The experiment is concerned with your ability to quickly skim the whole of or an extract of broadcast news stories. You will be given 15 questions, one at a time, and your task is to find in which news story the answer to the question can be found, i.e. I want you to return the* ***number*** *of the relevant news story. In order to make a quick judgement you will want to skim the text rather than read it thoroughly. Also note that you don't actually have to be able to answer the question, as long as you have located which story is likely to provide the answer. Please give your answer as soon as you're ready to make a qualified guess - don't worry too much about getting it wrong!*